AUTOMATED BATCH ARCHIVAL PROCESSING: PRESERVING ARNOLD WESKER'S DIGITAL MANUSCRIPTS

BY SARAH KIM, LORRAINE A. DONG, AND MEGAN DURDEN

ABSTRACT: Many cultural institutions and archives expect to acquire an increasingly copious amount of born-digital materials that will need to be processed at the item level. However, limitations in technology, labor, and funding resources often hinder small and medium-sized cultural institutions in their efforts to adequately implement and maintain long-term preservation methods for their digital archival holdings. The Arnold Wesker Digital Manuscript Project focused on the development of an automated batch archival processing work flow that minimized the labor and time required when handling large electronic archival collections. By working with various software and digital tools, it was possible to take advantage of affordable and accessible technology to successfully process more than five thousand digital files with a limited amount of time and resources. It is hoped that this automated batch archival processing approach will encourage other cultural institutions to initiate dynamic preservation projects for their digital collections.

Introduction

In the spring of 2007, three students in a graduate class focusing on the preservation of digital objects conducted a five-month project at the Harry Ransom Humanities Research Center (HRC) at The University of Texas at Austin under the supervision of Dr. Patricia Galloway, an associate professor in the School of Information at the university, and Catherine Stollar Peters, who was then an archivist at the HRC.[1] The primary goal of the project was to explore a long-term preservation method for digital manuscripts in current and future collections at the HRC.[2] The main project tasks included the processing and depositing of the HRC's digital manuscripts into the School of Information digital repository, supported by DSpace, an open source software package for managing digital assets.[3]

This hands-on experience with the HRC materials allowed the authors to perform, through trial and error, countless hours of rote effort, and a few epiphanies, the technical

and intellectual work necessary to engage in the archival processing of digital manuscripts. In the process, the team developed a work flow that involved "automated batch archival processing," a method that utilizes digital tools and relies on the machine-readable aspect of digital materials. Unlike paper records that require human eyes to read and process them, digital materials can be processed in groups by digital tools. For example, to create a file inventory of a hundred paper records, a human processor would need to go through the records one by one, where file-cataloging software could produce a file inventory of a hundred digital records within a few minutes. Although there are problematic issues in obtaining and using digital tools to meet individual collection needs, the authors were satisfied with the software and programs used in the project. It is possible to use these tools for every step of archival processing, from creating the file catalog to ingesting files into a digital repository.

For born-digital holdings, item-level processing is an especially important preservation activity for ensuring long-term accessibility of holdings. If individual disks and electronic files are not continuously supported by appropriate technical environments based on individual needs and future technical changes, technological obsolescence will greatly affect access. To extend an analogy from Mark Greene and Dennis Meissner's noteworthy work on minimum standards processing, archives will become "forests without any trees."[4] Thus, the automated batch archival processing can be regarded as a practical option for archivists processing digital holdings at the item level while also reducing the time and labor required for item-level processing in general. It will help archivists achieve the goal of "more product, less [human-involved] process."[5] Given further investigation, this approach has great potential to improve current digital preservation capabilities, especially for digital manuscripts in cultural institutions.

In this paper, the authors will share their experiences and findings, which archivists and preservation specialists may find helpful when considering whether batch processing of digital materials is a viable option for their collections.

Arnold Wesker Digital Manuscripts

The digital manuscripts that were the focus of this project were the personal digital materials of the British playwright Sir Arnold Wesker (1932–). Arnold Wesker is considered one of the key figures in twentieth-century drama. He has written 42 plays, as well as short stories, film and television scripts, poetry, essays, a children's book, a book on journalism, and an autobiography of his life since 1957. Wesker's plays have been translated into 17 languages and performed worldwide.[6]

The HRC acquired both the paper and digital manuscripts of Arnold Wesker on January 19, 2000, March 2, 2001, and May 20, 2003. The paper holdings, which were processed and stored in the HRC, contain a considerable amount of production materials for Wesker's plays as well as correspondence, hand-written drafts, and typescripts of Wesker's lesser-known works. These paper holdings are arranged in series identified as "Works and Related Materials," "Correspondence," "Personal," and "Works of Others."[7] The digital holdings of Arnold Wesker comprise electronic text files of correspondence, works, diaries, and other personal documents. Wesker used WordPerfect

5 and 9 and Microsoft Word 97 to create these text files. He organized them by subject and year, and saved them primarily on 3.5-inch floppy disks. The HRC received 75 3.5-inch floppy disks and one Zip disk from Wesker that contained more than seven thousand text files, for a total size of 100.01 MB.

Project Objective

The goal upon completion of archival processing was to have the individual files of the Wesker digital manuscript collection, which came to the HRC on obsolescent media, deposited in the institutional digital repository along with at least a minimal amount of metadata. Files would be arranged by provenance and in formats supported by the repository. Descriptive information such as the collection scope, content notes, and biographical sketch would also be provided. As Douglas Bicknese notes, preserving digital holdings in an institutional digital repository such as DSpace makes it easier for the archives to ensure that preserved files remain accessible, authentic, and reliable over time for the institution's designated user community.[8] This preservation method allows the institution to focus on managing the files on an individual basis in future preservation efforts rather than working with potentially nonviable media. Each item's associated metadata, which provides provenance and necessary technical information, will ensure its authenticity and long-term accessibility. Finally, a digital repository will encourage the maintenance of sustainable file formats.

Project Challenges and Approaches

While conducting a literature review of class reading materials, the authors developed a better understanding of how to work with digital repositories and the tasks involved in preparing and submitting files to an institutional repository. This review guided the overall project work flow.[9] While doing the actual processing, the group continually confronted unique challenges raised by both technical and archival preservation issues. In response to each challenge the authors sought appropriate approaches through research on digital tools and software, discussions with the School of Information IT staff, and a series of test runs.

Item-Level Processing vs. Quantity of Holdings

Since item-level processing was a major focus of the project, one of the biggest challenges for the group was the number of files. Although the total size of 100.01 MB appears manageable, the more than seven thousand electronic text files was a large number of files to process at the item level, given the limited amount of time and resources available. It was evident from the start of the project that it would be extremely labor intensive to process the electronic files manually one by one. As an alternate solution, the team sought out automated batch processing methods for every step of the project. This involved the testing and application of various file management software applications and customized programs such as Perl scripts.

File Extension in the WordPerfect Application

The WordPerfect application that Wesker used to produce most of his text files between 1989 and 1996 allowed the record creator to assign a three-character file extension instead of a standard file extension such as ".wpd" or ".doc." Wesker used this option in order to create a unique naming convention for his files that reflects the context and content of his documents. For example, the file name SHYLOCK.ACC conveys that this file is a document of bank account information for one of Wesker's plays, *Shylock*. In this case, ".ACC" is interpreted by a computer as a file extension, while Wesker intended it to be viewed as part of the file name. Since Wesker assigned these file extensions systematically, they contained valuable information and constituted an important property to preserve. However, these user-created file extensions also provided false technical information about the file type and made it difficult to render the files in the appropriate application. To address this, the team preserved the original files and in addition created access copies in Rich Text Format (RTF) with modified file names. Although RTF is a proprietary format, the team chose it due to its popularity and multiplatform capabilities; these features allow users to easily open and read the access copies of the Wesker holdings. As a result of the file conversion process, each access copy contains the original file extension as a part of its file name and ".rtf" as its file extension. For example, the original file called SHYLOCK.ACC was preserved along with its access copy in RTF format, called SHYLOCK.ACC.rtf. Further information about creating access copies will be described below.

Password-Protected Files

The Wesker digital collection contains 96 password-protected files that are diary entries. Although the HRC acquisition documentation for the Wesker collection contains the passwords for these files, it also states that these particular works are not to be made available to the public until 25 years after Wesker's death or after the death of his last surviving child, whichever comes last. Thus, the password-protected files are under a deep restriction policy and viewing them is limited to preservation purposes only. After discussion with Stollar Peters, the team decided that one of its members would take the responsibility of writing a preservation proposal for restricted access to these files and process them apart from the rest of the collection. They then deposited these password-protected files in DSpace for preservation purposes only and restricted access to HRC archivists.

Corrupted Disks and Files

The materials contained a few corrupted disks and files, which Wesker had noted on the labels of the disks. The group applied digital archeology techniques in an attempt to recover them, using a program called Dead Disk Doctor. It turned out this software was not an appropriate preservation choice because it alters the original bitstream. The team also asked IT professionals from the School of Information for technical advice. However, the team was not successful in extracting the information from the corrupted disks and files due to technical limitations. Team members recommended that the HRC keep the original disks in the hope that the technology to make them accessible may become available in the future.

Original Order

Prior to transferring his personal digital manuscripts to the HRC, Wesker organized the files by subject and year. The team considered this "internal method" of organizing files by the record creator to be the original order of the materials.[10] As one of the fundamental principles of archives, preservation of the original order is essential to show the context of individual files and to reflect Wesker's intentions. The question, however, was whether to maintain the original order of the electronic files or to rearrange the files to follow the existing arrangement scheme for the paper-based Wesker collection. If the latter practice were adopted, the corresponding relationship between the paper holdings and the digital holdings would become more evident. The authors decided to pursue this route, while at the same time capturing the original order of the digital manuscripts by documenting the original file directory structure, which was included in the file catalog and DSpace as one of the descriptive metadata elements.[11]

Duplicate Files

Wesker kept a number of backup files. These duplicate files raised two issues: how to make sure the files were exact duplicates and what to preserve among duplicate files. First, in order to assess whether or not the duplicate files were exact copies, the group ran the files through an Adler 32 checksum calculation and the Message-Digest algorithm 5 (MD5) digest calculation.[12] Files with the same checksum have identical bitstreams. Second, to save space on DSpace, the group decided to exclude duplicate files from the repository ingest.[13] However, duplicates could not be excluded without first considering their context within the file structure. For example, Wesker had saved two identical letters in two different subject folders—one organized by recipient and one organized by project—along with different files in each folder. In this case, even though the content of the two letters are the same, the context in which each letter is saved is different. After discussing these challenges with Stollar Peters, the authors decided to define specific criteria for deleting duplicate files; these criteria are described in the next section.

Archival Processing 1: File Processing

In the initial phase of the archival processing, the authors conducted a collection assessment and prepared files for long-term preservation. This included creating a disk inventory and a file catalog, creating access copies, excluding duplicate files for the repository deposit, and refreshing the files to a new medium.

Creating a Disk Inventory

The first step in processing was to inspect the physical materials and manually create a disk inventory. This initial inspection was critical to begin to understand the scope of the holdings and to determine the physical characteristics of the original materials that may be pertinent to future researchers. While examining the 75 3.5-inch floppy disks and one Zip disk, the team created a disk inventory using Microsoft Excel and devised these categories for the inventory:

- Disk type: the brand and model of disk
- Used size: amount of space used (in KB or MB)
- Description on disk: information written on the label of the disks
- Notes: information provided by Wesker that pertains to the files on the disks

Creating a File Catalog

The next step was to create a file catalog. With the exception of the corrupted disks, the folders and files from the original disks were copied and saved onto the hard drive of an HRC computer running Windows 2000. The team then used WinCatalog Light, a free cataloging software program, to create the file catalog.[14] This software automatically harvests certain types of metadata from individual files and presents the results in an Excel spreadsheet. It took approximately five minutes to create a catalog of seven thousand files with this software. The final categories included in the file catalog were:

- Name: file name
- File type: in general, file extensions define the file type; in this project, file types reflect Wesker's unique file-naming conventions
- Size: in MB and KB
- Last Modified Date: the last date and time on which the file was saved
- Location: on the disk; reflects directory structure that shows the original order in this project

After the group made the catalog, the Norton AntiVirus program checked it for viruses and the Jacksum program created checksums.[15]

Creating Access Copies

As mentioned above, the authors made access copies in RTF to increase accessibility of the original WordPerfect files with the unique file extensions.

As part of this project, the HRC purchased the ABC Amber Text Converter v. 4.10, a program that converts batches of one type of file to another.[16] Before using the file converter program, however, the team applied a customized Perl script to add the ".doc" files extension to the copies so that the original files were readable by the file converter program (e.g., SHYLOCK.ACC.doc).

Using the ABC Amber Text Converter v. 4.10, the group converted the DOC files to RTF files. Without the program, it would have been necessary to manually open each DOC file in WordPerfect 9 and save it as an RTF file, an overwhelming task given the number of files in the Wesker collection. Despite the automatic batch conversion, which saved a great deal of time and labor, all of the resulting RTF files had to be checked for accuracy, as some files did not convert for various reasons, such as password protection. These files were manually converted.

Excluding Duplicate Files

As mentioned above, the team excluded duplicate files from ingest into DSpace for space-saving purposes when they met certain criteria, which are outlined below. The group identified duplicates using the software zsDuplicateHunter Standard 2.31.[17] This program was set to use the Adler 32 checksum calculation and the MD5 message digest calculation to search for and display files. Although the program supports an option

to delete automatically all recognized duplicate files, it was used solely to identify duplicates; deletion involved an appraisal of the materials and a determination of how they reflect the creator's original organization structure.[18] The team manually deleted approximately two hundred files using the following criteria:

1. No duplicate files were deleted if they had any differing file and/or extension names.
2. No duplicate files were deleted if they were originally in two different subject folders.
3. A duplicate file was deleted if it was in one general folder such as a backup folder and its counterpart was in one specific subject folder. The authors reasoned that the remaining file carried more descriptive information than the one in the general folder.

Refreshing to New Media

The 3.5-inch floppy disks were in good physical condition at the time of acquisition and have been well maintained at the HRC. However, since the popularity of the 3.5-inch floppy disk as a medium is steadily declining, the group saved the original bitstreams onto CD. Access copies, the disk inventory, file catalogs, and the MD5 list created during processing were also saved on CD. The CDs are currently stored with the original 3.5-inch floppy disks at the HRC.

Archival Processing 2: Batch Ingesting into DSpace

The next phase of the archival processing was to deposit the processed files into DSpace. DSpace is an open source digital repository software developed by the Massachusetts Institute of Technology Libraries and Hewlett-Packard. More than two hundred institutions worldwide have used it to archive and store digital data in a variety of formats, from documents and spreadsheets to music and video.[19] At the School of Information, Dr. Galloway began experimenting with DSpace in 2002 and the School adopted it in 2005 as its permanent institutional digital repository.[20] Currently, the HRC uses the School of Information DSpace as its digital depository, planning eventually to install its own DSpace.

In DSpace, end-users can submit files one by one with appropriate descriptive metadata through a Web-based interface.[21] Single or multiple files packaged with necessary metadata are treated as an "item" and grouped in a "collection" in a DSpace "community." The basic structure of DSpace is as follows:

Community
Sub-community (bunches of collections and/or sub-communities)
Collection (bunches of items)
Item (bunches of bitstreams)
Bitstream

Since there were 12,000 bitstreams of original files and their access copies to deposit, the group employed a batch ingest method rather than using the Web-based manual file submission interface.[22] Batch ingest, or bulk uploading, is the use of specific command lines in the UNIX environment to load files and metadata directly into the DSpace database and file space. This method allows the DSpace administrator to submit as many files in one instance as needed.

It must be noted, however, that a batch ingest needs to be conducted in a specific manner in order to be compliant with the stringent DSpace requirements.[23] Files must be organized into both an acceptable directory structure and precisely named item folders. Furthermore, the files must be packaged with the appropriate metadata before the ingesting process begins. For this project, the batch ingest preparation tasks included creating metadata files, naming item folders under DSpace conventions, and creating contents text files for each of the item folders.

Metadata Preparation

The authors first harvested metadata from the bitstreams in the original format using the Metadata Extraction Tool, which was developed by the National Library of New Zealand.[24] This harvester automatically extracts metadata elements from multiple files assigned by the file's creating software, including the date of creation, date of modification, name of the record creator, title, and so forth. This tool outputs the extracted metadata in XML formats. Next, in order to abide by DSpace's current usage of the Dublin Core metadata schema, the team converted the extracted XML metadata files to Dublin Core metadata files by applying a modified Perl script.[25]

Completing the Batch Ingest Package

As mentioned above, the batch ingest process requires files to be packaged as an "item," which is a folder containing the materials necessary for submission and the metadata files (contents file and Dublin Core metadata file) required by DSpace. To prepare the items for ingest, various Perl scripts completed basic automatic batch tasks, such as the creation of empty folders, folder renaming, and the creation of contents files. The following example illustrates how an item ready for batch ingest into DSpace would appear:

item_001 (folder)
  contents: text file containing one line per filename (DSpace requirement)
  dublin_core.xml: Qualified Dublin Core metadata (DSpace requirement)
  SHYLOCK.ACC: bitstream in the original file format
  SHYLOCK.ACC.rtf: access copy
  SHYLOCK.ACC.xml: extra metadata file created by the New Zealand Metadata Harvester (optional)

Performing the Ingest

The group conducted a series of trial runs in order to identify errors prior to the actual batch ingest. As noted before, DSpace has very stringent file-naming and metadata requirements, some of which the team is still discovering. For example, during the trial

ingest process, characters such as "$," "£," "~," and "?" in file names were rejected by DSpace. To deal with this, the authors added a step at this point in the preparation to manually remove any potentially faulty characters from the file names and contents files. This could not have been done earlier in the process because metadata that reflected Wesker's original file names was needed. After two rounds of trials and correcting any errors, the group had ingested all of the items prepared. It took approximately three hours to ingest more than 5,500 items using a PC with Windows XP Professional in the School of Information computer lab.

Project Results and Suggestions

This project deposited a total of 5,757 items in four collections in the School of Information DSpace under the "Special Projects—Harry Ransom Humanities Research Center, Arnold Wesker Papers" sub-community. Each item contains the bitstream in its original file format, an access copy in RTF file format, and an extra metadata file. Although access to the content of individual files is currently restricted for use by HRC staff and on-site patrons only, the biographical sketch of Arnold Wesker, the sub-community and collection scope and content notes, brief metadata of each item, the file catalogs, and the project documentation reports are available to the public on-line.[26]

The Arnold Wesker digital sub-community and collection arrangement in DSpace is outlined below:

I. Works and Related Material: This series contains produced and unperformed works in Subseries A: Works by Title, including plays for stage, radio, television and screen, opera, ballet, musicals, short stories, and non-fiction. These works are arranged alphabetically.

IV. Correspondence: This series includes personal correspondence written by Wesker between 1989 and 1997, including communication with family, particularly his daughter and granddaughter; friends; actors; agents; and publishers. Correspondence related to his works can be found in Series I.

V. Personal: The documents in this series pertain to personal information such as finances, contracts, and writings by his mother.

VII. Restricted Materials: Files in this series may not be viewed by HRC patrons until the restrictions are removed. Items in this series include correspondence and diary entries.

Developing Automated Batch Archival Processing

The main achievement of this project is the development of a method that the authors call "automated batch archival processing," which utilizes digital tools to process

digital archival holdings in groups. The software and Perl scripts that the team tested and applied in this project reduced the time, human labor, and error in the item-level digital archival processing. However, the selection process for appropriate software was not always straightforward. The authors preferred open source software, but in many cases commercial software provided more options and user-friendly interfaces. Since the group could not find sufficient technical information to justify the purchase of certain software for archival purposes, they relied heavily on advertisements and testing results from trial versions of software. They discovered that Perl scripts have great potential to be used in a variety of digital archival processing tasks because they can be modified to meet the individual needs of collections or tasks. However, creating customized Perl scripts for a particular task required a certain level of computer programming knowledge and skill.

As a result of this project, the authors have determined that the development of automated archival processing toolkits is vital for the improvement of digital preservation capabilities, especially as vast quantities of digital holdings become more common in archival collections.

Automated batch archival processing toolkits should include the elements outlined in table 1.

Table 1: Recommended Elements for an Automated Batch Archival Processing Toolkit

Tool Type	Purpose	Items and Function *(tools applied in this project)*
Collection assessment	To obtain overall characteristics of holdings, such as size, file formats, location, and modified/created dates	• File-cataloging tool: Automatically create file catalog (WinCatalog Light) • File format–identifying tools: Identify unknown file formats (*DROID and JHOVE)*
File preparation	To prepare files for long-term preservation or deposit in an institutional repository	• Virus-checking tool: Check virus and malware *(Norton AntiVirus)* • Redundancy-checking and file-integrity-checking tool *(Jacksum)* • Duplicate-identification tools: Identify technically identical files *(zsDuplicateHunter Standard 2.31)* • Batch file-conversion tool: Create access copies in or convert files to more sustainable file format *(ABC Amber Text Converter v. 4.10)* • Multiversion-control tool: Treat multiversion files
Metadata preparation	To harvest metadata; To create/edit metadata especially for facilitating crosswalks between different metadata schemas	• Metadata-harvesting tool: Automatically extract metadata from files *(New Zealand Metadata Harvester)* • Metadata-editing tool: Edit/add metadata in a desired form or schema *(Perl scripts to convert extracted metadata into Dublin Core XML form)*
Digital archeology	To recover and/or extract bitstreams from damaged media or corrupted files	• Tools may vary from a physical media-cleaning tool to an image-mapping tool or bitstream-level recovery tool that does not alter original bitstream.
Item management	To manage folders and/or files in groups effectively	• Tools will vary according to the task, such as batch folder/file creation, renaming, directory changing, deletion, and so forth. *(Customized Perl scripts for batch folder/file creation and renaming)*
Step-by-step guidelines	To provide guidance for appropriate installation and usage of software for nonexperts; To provide guidance for proper handling of digital holdings	• Textual/visual tutorials

In order to build an automated archival-processing toolkit, open source software and programs need to be designed to reflect the specific needs of digital archives. It should be kept in mind, though, that ready-made software provided by vendors for general digital file management can be included as digital archival processing tools. The path to creating a successful toolkit will require long-term investigation supported by institutional interests and stable funding. Furthermore, archivists and record managers must collaborate with software designers to produce robust archival tools.

Knowledge Sharing

The work associated with managing outdated file formats, password-protected items, and a large number of files is labor-intensive and time-consuming without proper technical knowledge. This project, however, was not the first to confront these challenges. The Presidential Electronic Records Project, sponsored by the National Archives and Records Administration in 1999, encountered similar problems with WordPerfect document files and password-protected objects.[27] Although the approach from the current project was quite different from the solutions of the Presidential Electronic Records Project, experience gained from both provided valuable insight. Other predecessors include the Cornell University Library DSpace, which provides on-line batch ingest guidelines for DSpace administrators to preserve their theses and dissertations collection,[28] and the Queen's University QSpace bulk upload standard operating procedures, which contain detailed information about how to bulk upload, including how to prepare Dublin Core metadata files for batch ingest.[29] Finally, the Michel Joyce Collection Digital Preservation Project, conducted by Stollar Peters and her colleagues as students in the same class for the HRC in 2005, served as a reference model for this project.[30]

When archivists work with personal records in digital form, the issues of formats and the technical environment of file creation become complicated. As Tom Hyry and Rachel Onuf point out, individual authors use many different computing systems and software to create their personal documents and, unlike in organizational settings, a standardized or universally shared format does not exist.[31] As a result, archivists expect a wide variety of digital manuscripts in terms of formats, media, and applications necessary to read the bitstreams. Sharing experiences about various types of digital manuscripts among archival institutions is crucial for determining suitable preservation methods for each unique personal digital collection. Building a centralized information pool for object-oriented digital archival preservation methods is vital.

Working with the Record Creator

As indicated above, the dynamic nature of digital manuscripts is a reminder of the importance of the record creator's role in archival processing. The record creator is the most knowledgeable person about the characteristics of items in a collection. Descriptive and technical metadata provided by a record creator is helpful at the beginning of archival processing. When Wesker transferred his files to the HRC, he made a list of the contents of each disk. One of the notable characteristics of electronic media is that it is impossible to know the types of electronic files contained in a given disk until the disk is actually examined using a non-destructive digital tool. Therefore, the

descriptive metadata provided by Wesker was highly useful as a starting point in the collection-assessment process. Wesker also provided short notes about the software that he had used to create the files. Although this information was neither complete nor consistent over the entire collection, the group relied on this technical metadata in order to examine files using the appropriate vendor software.

Archivists and records managers have emphasized the important role of record creators in the long-term preservation of government or business electronic records. This notion is also true for personal digital manuscripts. Adrian Cunningham suggests what he calls "pre-custodial intervention," in that archivists work with active record creators early in their careers to ensure that personal electronic records are properly created, managed, and documented in the first instance, thus improving the ability to preserve and provide access to these records in the long run.[32] Even if archivists are accepting personal digital materials from donors who are near the end of their lives or careers, collaboration with living record creators is still critical and should begin in the collection-acquisition stage. If possible, digital archivists should gather essential information about the record creator's digital working environment, which includes operating systems, applications, personal file organization scheme, file-naming method, and so forth. This information is vital for deciding the appropriate preservation approaches for each collection.

Conclusion

Management and preservation methods for electronic records have been developed and implemented through various electronic record preservation projects, especially in business and organizational settings. A number of commercial sectors offer digital archiving services for institutions. However, it is relatively difficult for small or medium-sized cultural institutions to create their own strategy for long-term preservation of their digital archival holdings. The unpredictable nature of digital manuscripts collected by cultural institutions tends to make them more unwieldy than organizational electronic records. Furthermore, heritage preservation institutions must face limited technology, labor, and funding resources. Although the Harry Ransom Humanities Research Center has a number of resources, especially in terms of funding, IT staff, and student volunteers, the tools the authors used to implement the batch processing of the Wesker collection were either freeware or available for a low cost. It is the authors' hope that this experience with digital manuscripts and the explicit instructions provided in the project documentation will encourage archivists in small or medium-sized cultural institutions to try the automated batch archival processing work flow in order to more efficiently preserve digital manuscript holdings.

Many small and medium-sized archives may not yet have an on-line repository such as DSpace set up. The goal of automated archival processing, however, is to prepare digital files for long-term access and preservation. The resulting formats and storage media from such batch processing are varied and dependent on each institution's need and capabilities. Institutions can use external storage devices and a network server as affordable storage. An on-line repository nevertheless remains the more sustainable

option to keep preserved materials accessible, reliable, and authentic. This project's final report, with detailed information about the processing work flow, the batch ingest manual, and the Perl scripts used in this project, is available through the School of Information DSpace at The University of Texas at Austin.[33]

***ABOUT THE AUTHORS*:** Sarah Kim is a doctoral student specializing in digital preservation in the School of Information at The University of Texas at Austin. She received her M.S. in Information Studies from the University at Albany, the State University of New York, in 2005.

Lorraine A. Dong is a doctoral student in the School of Information at The University of Texas at Austin. She received her M.S. in Information Studies and a Certificate of Advanced Study in Preservation Administration from the same institution in 2008, and her M.Phil. in Renaissance literature from Cambridge University in 2005.

Megan Durden is an archivist in the Electronic Records Unit at the State Archives of North Carolina. She received her M.S. in information studies from the School of Information at The University of Texas at Austin in 2007 and a Certificate of Advanced Study in Preservation Administration of Library and Archival Materials from the Kilgarlin Center for Preservation of the Cultural Record.

The authors wish to thank the following people for their advice, support, and participation: Dr. Patricia Galloway, Catherine Stollar Peters, Shane Williams, and Sam Burns.

NOTES

1. In the spring of 2007 the authors enrolled in INF392K, Problems in Permanent Retention of Electronic Records, instructed by Dr. Patricia Galloway in the School of Information, The University of Texas at Austin.
2. "Manuscript" here is a term adopted from the Society of American Archivist's "A Glossary of Archival and Records Terminology" to describe unpublished material that is created or gathered by an organization or an individual. "Digital manuscript" is used to distinguish a born-digital manuscript from a traditional handwritten or typewritten manuscript.
3. Currently, the HRC uses the School of Information DSpace as its digital depository, planning eventually to install its own DSpace
4. In their intensive review of archival literature, Greene and Meissner mentioned that there is substantial evidence in the archival profession to support taking a "forest-not-trees" approach to processing that emphasizes series-level description rather than item- or folder-level processing, so as to maximize the availability of holdings. However, for born-digital holdings, item-level processing and maintenance is vital to insure their accessibility in spite of the rapid change of technology. Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *American Archivist* 68:2 (2005): 208–263.
5. Greene and Meissner, ibid., 208–263.
6. Arnold Wesker, "Homepage," http://www.arnoldwesker.com (9 April 2007); Cathy Grindrod, "Arnold Wesker," *Contemporary Writers*, http://www.contemporarywriters.com/authors/?p=auth224 (9 April 2007); Anne Etienne, "Arnold Wesker," *Literary Encyclopedia*, December 7, 2004, http://www.litencyc.com/php/speople.php?rec=true&UID=4663 (9 April 2007).
7. The finding aid for the HRC Arnold Wesker paper collection can be found at http://research.hrc.utexas.edu:8080/hrcxtf/view?docId=ead/00243.xml. However, the current finding aid does not contain the information about Wesker's digital holdings.

8. Douglas Bicknese, "Institutional Repositories and the Institution's Repository: What Is the Role of University Archives with an Institution's On-line Digital Repository?" *Archival Issues* 28:2 (2003-2004): 90.
9. Among the class reading materials, the following articles were particularly useful for this project: US-InterPARES Project, "Findings on the Preservation of Authentic Electronic Records," September 2002, http://www.gseis.ucla.edu/us-interpares/pdf/interpares1finalreport.pdf (5 February 2007); ERPANET, "Erpa Guidance: Ingest Strategies," September 2004, http://www.erpanet.org/guidance/docs/ERPANETIngestTool.pdf (20 March 2007); Consultative Committee for Space Data Systems, "Producer-Archive Interface Methodology Abstract Standard," May 2004, http://public.ccsds.org/publications/archive/651x0b1.pdf (20 March 2007); William G. LeFurgy, "Levels of Service for Digital Repositories," *D-Lib Magazine* 8:6 (May 2002). The entire list of reading materials assigned to the class is available at the 2007 class Web site, http://courses.ischool.utexas.edu/galloway/2007/spring/INF392K/schedule.html.
10. Tom Hyry and Rachel Onuf, "The Personality of Electronic Records: The Impact of New Information Technology on Personal Papers," *Archival Issues* 22:1 (1997): 39.
11. The file catalog contains the location information for each file. For example the location information of the file AUKIN.SCR is WESKER\R14620.2-49\BLOODLIB\COR.92\. The last two directories of this file location indicate that this file was originally organized and saved under the folder "BLOODLIB" and its subfolder "COR.92," by Wesker. All folders and subfolders were created and named by Wesker. This original file directory information was also included in DSpace as a "description" metadata element with a "URI" qualifier.
12. Adler 32 checksum is a form of redundancy check using 32-bit temporary sums. Message-Digest algorithm 5 (MD5) is used for message integrity checks and to perform digital signatures for a data stream that represents the contents of digital documents.
13. The Open Archival Information System (OAIS) reference model uses "ingest" to mean processes related to accepting submitted information packages from an external source and preparing them as archival information packages for storage. Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," January 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf (1 February 2007).
14. The team tested several types of file cataloging software and ultimately chose WinCatalog Light because it is freeware and is relatively easy to install and use.
15. Jacksum is a platform-independent checksum utility.
16. ABC Amber Text Converter was selected because the price was affordable and the software handles diverse file formats.
17. ZsDuplicateHunter is available for a reasonable price.
18. Determining the intellectual value of the individual files was beyond the scope of this project. Therefore, appraisal extended only to removing duplicates for ingest.
19. More information about DSpace can be found at the DSpace official Web site, http://www.dspace.org/.
20. DSpace at the School of Information, The University of Texas at Austin, can be found at https://pacer.ischool.utexas.edu/.
21. DSpace currently supports the Dublin Core metadata schema based on the Dublin Core Libraries Working Group Application Profile.
22. The School of Information DSpace Batch Ingest Guide prepared by Sarah Kim, Lorraine A. Dong, and Patricia Galloway can be found at https://pacer.ischool.utexas.edu/handle/2081/9228.
23. DSpace, "DSpace System Documentation: Application Layer," http://www.dspace.org/index.php?option=com_content&task=view&id=144 (15 March 2008).
24. More information about this metadata extraction tool can be found at http://meta-extractor.sourceforge.net/.
25. Stollar Peters previously tested this particular Perl script when processing the Michael Joyce digital collection at the HRC. Several elements of this Perl script were modified for the Wesker digital collection.
26. The Harry Ransom Humanities Research Center, Arnold Wesker Papers, can be found at https://pacer.ischool.utexas.edu/handle/2081/2220.
27. William E. Underwood, "Analysis of Presidential Electronic Records: Final Report," September 1999, http://perpos.gtri.gatech.edu/perpos/Final_Report.pdf (24 February 2007).

28. Holly Mistlebauer and George Kozak, "DSpace Administration," *Cornell University Library's Wiki*, http://wiki.library.cornell.edu/wiki/x/0S8 (15 April 2007). This Web page is no longer available.
29. Seamus Ryan, "Queen's University QSpace: Bulk upload standard operating procedures," April 11, 2005, http://library.queensu.ca/webir/qspace-project/tutorials/qspace_bulk_upload.doc (2 May 2007).
30. Catherine Stollar Peters, "When Not All Papers Are Papers: A Case Study in Digital Archivy," *Provenance* 24:1 (2006): 23–35.
31. Hyry and Onuf, 37–44.
32. Adrian Cunningham, "Waiting for the Ghost Train: Strategies for Managing Personal Electronic Records before It Is Too Late," *Archival Issues* 24:1 (1999): 55–64.
33. The project documentation is available at https://pacer.ischool.utexas.edu/handle/2081/2322.